NEAR-FIELD ADAPTIVE BEAMFORMING AND SOURCE LOCALIZATION IN THE SPACETIME FREQUENCY DOMAIN

Francisco Pinto and Martin Vetterli

Ecole Polytechnique Fédérale de Lausanne
EPFL-IC-LCAV, Station 14, CH-1015 Lausanne, Switzerland
francisco.pinto@epfl.ch; martin.vetterli@epfl.ch

ABSTRACT

We revisit the topics of near-field adaptive beamforming and source localization following an alternative approach based on a spatio-temporal spectral representation of the acoustic wave field. With the proposed method, the wave field is expressed as a separable combination of the signal and spatial components that characterize the various sources in the acoustic scene. This allows beamforming operations such as beam steering and sidelobe canceling to be translated into a two-dimensional (2D) sampling problem, where the sampling kernels are derived according to a parametric model representing the 2D spectral pattern generated in the presence of a source. Conversely, the spectral pattern can be estimated from an arbitrary input through the use of parametric spectral estimation techniques, providing a novel solution to the near-field source localization problem.

***Index Terms*—** Adaptive beamforming, sidelobe canceling, source localization, spatio-temporal processing, spectral estimation.

1. INTRODUCTION

In array signal processing, adaptive beamforming is a technique for directional sound acquisition that typically uses its own input or output signals to adjust the parameters of the system. Beamformers of this type have been studied, for example, by Frost [1] and Widrow *et al* [2]. In particular, a technique introduced by Griffiths *et al* [3], known as sidelobe canceling, is closely related to the method discussed in this paper. The idea consists of steering the zeros of the directivity pattern towards the target direction, such that only the background noise is captured. This noise is then canceled out in the beamformer output through adaptive filtering, in order to maximize the output signal-to-noise ratio (SNR).

In this paper, we propose a near-field beamforming approach (see Fig. 1) conceptually similar to sidelobe canceling except that the processing is performed in what we call the spacetime frequency domain [4], obtained by taking the two-dimensional Fourier transform along the temporal dimension and the spatial dimension (representing the array axis) of the multichannel input. In this domain, the sound acquisition can be directed towards (or null out) a given point in space by performing a 2D sampling operation, where the sampling indexes are determined by the expected spectral pattern generated by a source located at the same point. Such a pattern, as we show, is a product of a *signal component* given by the source signal and a *spatial component* given by the source location. In this context, the concept of beamforming is equivalent to estimating the signal component in the observed spectral pattern.

This project is funded by the *Fundação para a Ciência e Tecnologia* (Portugal) and the *National Center of Competence in Research for Mobile Information and Communication Systems* (Switzerland).

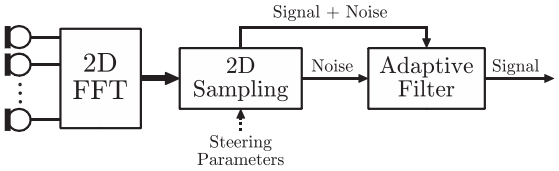

Fig. 1. Adaptive near-field beamformer based on spatio-temporal spectral sampling of the input multichannel data. The beam-steering parameters of the beamformer (angle and distance) can be estimated from the input spectra.

Equivalently, the spatial component can be estimated in order to localize the point source in space, with the main difference that there are only two parameters to estimate: the angle and the distance. In this paper, we show how the two parameters can be estimated through the use of parametric estimation techniques - in particular, template matching. A simulation result with a speech source in the near-field and a white noise interferer in the far-field is provided as an example.

2. ACOUSTICAL MOTIVATION

2.1. Spacetime spectral analysis

Consider the two acoustic scenes depicted in Fig. 2. A point source in free-field is typically characterized by a source signal $s(t)$ and a spatial position $\mathbf{r}_o = (x_o, y_o)$, assuming that the source is located on the $z = 0$ plane. These two parameters are enough to obtain the sound pressure $p(x,t)$ at any point along the x-axis, representing the microphone array. The result (without fixed amplitude factors) is given by [5]

$$p(x,t) = \frac{1}{\|x - \mathbf{r}_o\|} s\left(t - \frac{\|x - \mathbf{r}_o\|}{c}\right), \qquad (1)$$

where c is the speed of sound and $||.||$ is the regular l_2-norm. The point source can be either in the near-field (NF) or the far-field (FF) depending on its distance to the point of observation. Denoting the 2D Fourier transform of (1) as $P(\Phi,\Omega) = \mathcal{F}_{x,t}\{p(x,t)\} = \int\int_{-\infty}^{\infty} p(x,t)e^{-j(\Phi x+\Omega t)}dtdx$, the results for the near-field and far-field cases are given by [6]

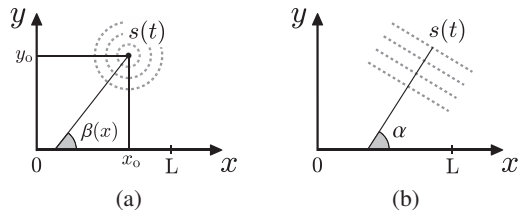

Fig. 2. (a) Near-field and (b) far-field acoustic scenes. The source is located either at a close range with coordinates (x_o, y_o) or very far away with a relative angle α. The array is defined on the x-axis from 0 to L, and $\beta(x)$ is the near-field angle of incidence at x.

$$P_{\text{nf}}(\Phi,\Omega) = S(\Omega)H_o^{(1)*}\left(y_o\sqrt{\left(\frac{\Omega}{c}\right)^2-\Phi^2}\right)e^{-jx_o\Phi} \quad (2)$$

$$P_{\text{ff}}(\Phi,\Omega) = S(\Omega)2\pi\delta\left(\Phi-\cos\alpha\frac{\Omega}{c}\right), \quad (3)$$

where Φ and Ω are the spatial and temporal frequencies, $S(\Omega)$ is the Fourier transform of $s(t)$, and $H_o^{(1)}$ is the zeroth-order Hankel function of the first kind. The result in (2) represents a spectrum of triangular shape where most of the energy is distributed across the region $|\Phi| \leq \left|\frac{\Omega}{c}\right|$, whereas in (3) the whole energy is concentrated in a single Dirac line of slope $\frac{\cos\alpha}{c}$ [6].

2.2. Windowing effects

Consider a window function $w(x)$ applied to the spatial dimension, such that $P(\Phi,\Omega) = \mathcal{F}_{x,t}\{w(x)p(x,t)\}$. Under the far-field assumption expressed in (3), it is simple to show that

$$P_{\text{ff}}(\Phi,\Omega) = S(\Omega)W\left(\Phi-\cos\alpha\frac{\Omega}{c}\right), \quad (4)$$

where $W(\Phi)$ is the Fourier transform of the window function (with unit amplitude). In particular, if $w(x)$ is a rectangular window of length L, the result is given by

$$P_{\text{ff}}(\Phi,\Omega) = S(\Omega)L\text{sinc}\left(\frac{L}{2\pi}\left(\Phi-\cos\alpha\frac{\Omega}{c}\right)\right), \quad (5)$$

which is illustrated in Fig. 3-a for $S(\Omega) = 1$. This result converges to the ideal solution in (3) as L tends to infinity, since $\lim_{L\to\infty} L\text{sinc}\left(\frac{L}{2\pi}\left(\Phi-\cos\alpha\frac{\Omega}{c}\right)\right) = 2\pi\delta\left(\Phi-\cos\alpha\frac{\Omega}{c}\right)$.

An important consequence of using the rectangular window, or other windows of the same class, is that the resulting two-dimensional spectrum has periodic zeros at $\Phi = \cos\alpha\frac{\Omega}{c} + m\frac{2\pi^2}{L}$, for every non-zero integer m. On the other hand, the weight is maximum at $\Phi = \cos\alpha\frac{\Omega}{c}$, for which $P(\Phi,\Omega) = S(\Omega)$.

In the near-field case, the result involves a convolution between (2) and $W(\Phi)$, which generates a spectral pattern as depicted in Fig. 3-b. Although this pattern is difficult to express mathematically, it can be intuitively understood as combination of two effects: (i) an orientation towards the source, as in Fig. 3-a, and (ii) a triangular spreading of the energy caused by the proximity of the source. With this in mind, we define a parametric model for the near-field spectrum as follows.

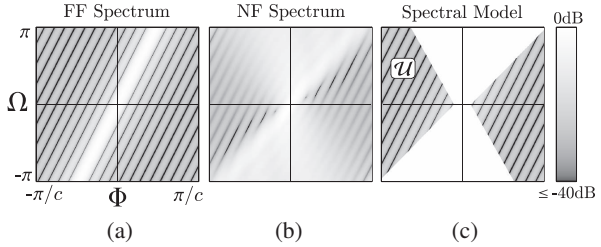

Fig. 3. Spectral pattern on the windowed x-axis generated by: (a) a far-field source with $\alpha = \frac{\pi}{4}$ and (b) a near-field source with the same spectral orientation. The respective parametric spectral model is shown in (c), where $\mathcal{U}$ is the region that resembles the far-field spectral pattern.

$$P_{\text{nf}}(\Phi,\Omega) \approx S(\Omega)\max\left\{W\left(\Phi-\cos\alpha\frac{\Omega}{c}\right), M(\Phi,\Omega)\right\}, \quad (6)$$

where $M(\Phi,\Omega)$ is a triangular mask given by

$$M(\Phi,\Omega) = \begin{cases} 1 & , (\Phi,\Omega) \notin \mathcal{U} \\ 0 & , (\Phi,\Omega) \in \mathcal{U} \end{cases}, \quad (7)$$

with $\mathcal{U} = \mathbb{R}^2 \backslash \left\{(\Phi,\Omega) : \cos\beta(L)\frac{\Omega}{c} \leq \Phi \leq \cos\beta(0)\frac{\Omega}{c},\ \Omega \geq 0\right\}$ (the limits swap for $\Omega < 0$), and $\cos\alpha = \mathbb{E}_x[\cos\beta(x)]$ is the spectral orientation, where $\mathbb{E}_x$ denotes expectation over x. This model is illustrated in Fig. 3-c for $S(\Omega) = 1$. Note that the result in (6) converges to (4) as the source moves away from the array line, given that $M(\Phi,\Omega)$ gets narrower as $\beta(x)$ tends to α.

2.3. The estimation problem: beamforming *vs* localization

A common aspect in the cases analyzed above is that the wave field is represented as a product of a signal component $S(\Omega)$ and a spatial component $B(\Phi,\Omega)$, such that $P(\Phi,\Omega) = S(\Omega)B(\Phi,\Omega)$, where the components are independent of each other. In particular, (6) can be used to obtain the other spectral patterns, either by setting $||\mathbf{r}_o|| \to 0$, $||\mathbf{r}_o|| \to \infty$, or $L \to \infty$.

In a scene composed of multiple sources, the superposition principle implies that

$$P(\Phi,\Omega) = S_0(\Omega)B_0(\Phi,\Omega) + S_1(\Omega)B_1(\Phi,\Omega) + \cdots \quad (8a)$$

$$= \sum_l S_l(\Omega)B_l(\Phi,\Omega), \quad (8b)$$

where $S_l(\Omega)$ and $B_l(\Phi,\Omega)$ are the signal and spatial components of each source in the acoustic scene. In this framework, the difference between beamforming and localization of a source l is related to the difference between estimating S_l and B_l, or a combination of both. For instance, if the target source is $l = 0$, the estimation can be performed in the following combinations: S_0 (beamforming); S_0B_0 (beamforming with preserved spatial cues); B_0 (localization). The main difference is that, while S_0 is generally a non-parametric signal, B_0 is completely defined by only two parameters: the angle and distance to the array. Thus, in the spacetime frequency domain, source localization is a parametric estimation problem.

In the next two sections, we present the problems of near-field beamforming and localization from the perspective of spatio-temporal processing, and show how these can be translated into a sampling problem and a parametric estimation problem.

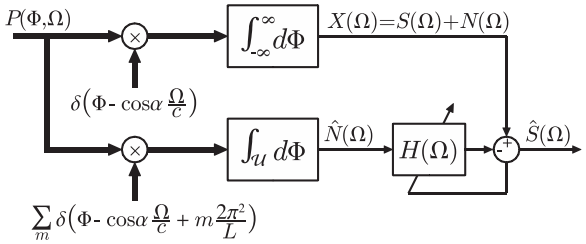

Fig. 4. Spectral sampling at the main lobe location (upper branch) and the zeros (lower branch) of the sinc support function. The estimated noise $\hat{N}(\Omega)$ is removed from $\hat{X}(\Omega)$ through adaptive filtering.

3. BEAMFORMING (SIGNAL ESTIMATION)

A closer look to the spectral pattern of Fig. 3-c shows that the signal component $S(\Omega)$ is differently weighted across the spectrum: while outside the region $\mathcal{U}$ the weight is maximum, within the region $\mathcal{U}$ the weight oscillates with periodic zeros. At the location of these zeros is where the background noise is expected to have higher energy than the target signal. The background noise can thus be estimated by sampling the spectrum over the parallel lines defined by $\Phi = \cos\alpha\frac{\Omega}{c} + m\frac{2\pi^2}{L}$ within the region $\mathcal{U}$, and subsequently canceled out from the noisy signal using adaptive filtering. This method, which we explain next in detail, is illustrated in the block diagram of Fig. 4.

Consider the following model of the acoustic scene, where a target source is mixed up with K noise sources at different locations in space, such that

$$P(\Phi,\Omega) = S(\Omega)B(\Phi,\Omega) + \sum_{k=0}^{K-1} N_k(\Omega)B_k(\Phi,\Omega) \tag{9}$$

where $N_k(\Omega)$ is the source signal of each interferer and $B_k(\Phi,\Omega)$ the respective spatial components. Consider also that the target source is in the near-field and the noise sources are in the far-field (which also models near-field sources, provided that K is large enough [5]), such that

$$P(\Phi,\Omega) = S(\Omega)\max\left\{W\left(\Phi - \cos\alpha\frac{\Omega}{c}\right), M(\Phi,\Omega)\right\} + \sum_{k=0}^{K-1} N_k(\Omega)W\left(\Phi - \cos\alpha_k\frac{\Omega}{c}\right), \tag{10}$$

where $\cos\alpha$ is the spectral orientation and $\alpha_k \neq \alpha$. The sampling kernels are then defined as

$$\Delta_S(\Phi,\Omega) = \delta\left(\Phi - \cos\alpha\frac{\Omega}{c}\right) \tag{11}$$

$$\Delta_N(\Phi,\Omega) = \sum_m \delta\left(\Phi - \cos\alpha\frac{\Omega}{c} + m\frac{2\pi^2}{L}\right), \tag{12}$$

where Δ_S and Δ_N denote *signal kernel* and *noise kernel* respectively. The signal kernel is used to obtain the spectral profile across $\Phi = \cos\alpha\frac{\Omega}{c}$, such that

Number of mics	**2**	**4**	**8**	**16**	**32**
Distance r (cm)	1.25	2.5	5	10	20
Attenuation (dB)					
$(\Phi,\Omega) \in \mathbb{R}^2$	8	11	11	11	11
$(\Phi,\Omega) \in \mathcal{U}$	8	14	20	25	30

Table 1. Sidelobe attenuation in the presence of a near-field source, where $r = \left\|\frac{L}{2} - \mathbf{r}_o\right\|$ is the distance to $x = \frac{L}{2}$. In all cases, the combination of values results in the spectral structure of Fig. 3-b.

$$X(\Omega) = \int_{-\infty}^{\infty} \Delta_S(\Phi,\Omega)P(\Phi,\Omega)d\Phi \tag{13a}$$

$$= S(\Omega) + \sum_{k=0}^{K-1} N_k(\Omega)W\left(\gamma_k\frac{\Omega}{c}\right) \tag{13b}$$

$$= S(\Omega) + N(\Omega), \tag{13c}$$

where $X(\Omega)$ corresponds to a noisy version of the signal $S(\Omega)$, and $N(\Omega) = \sum_{k=0}^{K-1} N_k(\Omega)W\left(\gamma_k\frac{\Omega}{c}\right)$ is the noise signal to be estimated and canceled. The factor γ_k indicates the level of separation between the target source and the noise sources, and is given by

$$\gamma_k = \cos\alpha - \cos\alpha_k = \begin{cases} 2 & , |\alpha - \alpha_k| = \pi \\ 0 & , \alpha = \alpha_k \end{cases}. \tag{14}$$

Assuming that $w(x)$ is rectangular, the two values displayed in (14) represent a worst-case and a best-case scenario, in the sense that: when the target source is on a coincident line with the noise sources, the argument of $W(\gamma_k\frac{\Omega}{c})$ is smaller and the noise energy increases in $X(\Omega)$, whereas when the signal and noise sources are totally separated, the argument of $W(\gamma_k\frac{\Omega}{c})$ is larger and the noise energy decreases.

The estimation of $N(\Omega)$ is obtained with the sampling kernel defined in (12), which is designed to sample the spectrum at the locations where $B(\Phi,\Omega)$ is zero, in accordance to (6). Therefore, special care must be taken not to sample $P(\Phi,\Omega)$ outside the region $\mathcal{U}$, where $B(\Phi,\Omega)$ has non-zero energy due to the source proximity effect. This also means that, for certain cases, there may be frequencies at which it is not possible to obtain an estimation of the background noise - typically at higher frequencies. Otherwise, the estimation of $N(\Omega)$ is given by

$$\hat{N}(\Omega) = \int_{\mathcal{U}} \Delta_N(\Phi,\Omega)P(\Phi,\Omega)d\Phi \tag{15a}$$

$$= \sum_{k=0}^{K-1} N_k(\Omega) \sum_{m\in\mathbb{Z}_{\mathcal{U}}} W\left(\gamma_k\frac{\Omega}{c} + m\frac{2\pi^2}{L}\right), \tag{15b}$$

where $\mathbb{Z}_{\mathcal{U}} = \left\{m : \left(\gamma_k\frac{\Omega}{c} + m\frac{2\pi^2}{L}, \Omega\right) \in \mathcal{U}\right\}$ for a non-zero integer m. Once $\hat{N}(\Omega)$ has been obtained, the denoising of $X(\Omega)$ becomes a Wiener filtering problem, which can be solved through adaptive filtering [3]. Ideally, this requires that $\hat{N}(\Omega)$ is maximally correlated with $N(\Omega)$ and uncorrelated with $S(\Omega)$, which only occurs in the far-field case. In the near-field case, the profiles taken at $\Phi = \cos\alpha\frac{\Omega}{c} + m\frac{2\pi^2}{L}$ are not exactly zero-valued, due to the energy spreading that affects the entire spectrum (see Fig. 3-b). Instead, there is a limited attenuation that can be obtained as long as the profiles are taken within the region $\mathcal{U}$ as opposed to $\mathbb{R}^2$. This is exemplified in Table 1.

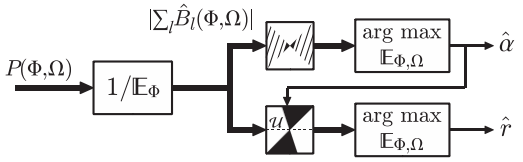

Fig. 5. Parameter estimation through template matching. The sum of spatial components in the wave field is estimated from $P(\Phi, \Omega)$ with a non-linearity. The result is then correlated with the angular and radial template functions in order to estimate the location of the sources.

4. LOCALIZATION (SPATIAL ESTIMATION)

The beam-steering parameters displayed in Fig. 1 can be either specified by the user or estimated from the multichannel input. In the spacetime frequency domain, estimating the location of the sources is equivalent to estimating the spatial components $B_l(\Phi, \Omega)$ in (8b). Since there are only two parameters in each component (angle and distance), the estimation can be done through the use of parametric estimation with an input model given by $\sum_l B_l(\Phi, \Omega)$ plus some noise component. If the sources are in the far-field, $B(\Phi, \Omega)$ has a nearly sinusoidal structure (Fig. 3-a) which allows the use of powerful techniques such as the Annihilating Filter and the MUSIC algorithm. In the near-field, this requires a more complex approach that takes into account the triangular spreading of the energy (Fig. 3-b). In this paper, we propose a technique based on template matching.

The spatial components in (8b) can be estimated by canceling out the signal components with a non-linearity,

$$|\sum_l \hat{B}_l(\Phi, \Omega)| = \frac{|P(\Phi, \Omega)|}{\mathbb{E}_\Phi\left[|P(\Phi, \Omega)|\right]}, \tag{16}$$

where $\mathbb{E}_\Phi$ denotes expectation over Φ. The parameters α and r are then separately estimated by correlating $|\sum_l \hat{B}_l(\Phi, \Omega)|$ with the template functions $T_\alpha(\Phi, \Omega)$ and $T_r(\Phi, \Omega)$, such that

$$\hat{\alpha} = \arg\max_\alpha \mathbb{E}_{\Phi,\Omega}\left[|\sum_l \hat{B}_l(\Phi, \Omega)| T_\alpha(\Phi, \Omega)\right] \tag{17}$$

$$\hat{r} = \arg\max_r \mathbb{E}_{\Phi,\Omega}\left[|\sum_l \hat{B}_l(\Phi, \Omega)| T_r(\Phi, \Omega)\right], \tag{18}$$

where the normalized templates functions are defined as

$$T_\alpha(\Phi, \Omega) = \left.\frac{W\left(\Phi - \cos\alpha\frac{\Omega}{c}\right)^l \overline{M}(\Phi, \Omega)}{\sqrt{\mathbb{E}_{\Phi,\Omega}\left[W\left(\Phi - \cos\alpha\frac{\Omega}{c}\right)^{2l} \overline{M}(\Phi, \Omega)^2\right]}}\right|_{r=0} \tag{19}$$

$$T_r(\Phi, \Omega) = \left.\frac{M(\Phi, \Omega)}{\sqrt{\mathbb{E}_{\Phi,\Omega}\left[M(\Phi, \Omega)^2\right]}}\right|_{\alpha=\hat{\alpha}}. \tag{20}$$

The inverse spectral mask is given by $\overline{M}(\Phi, \Omega) = 1 - M(\Phi, \Omega)$, and the parameter l controls the attenuation of the sidelobes of the sinc function. The angular template function assumes a worst-case scenario with maximum openness of the triangular mask ($r = 0$), whereas the radial template function takes advantage of the angle estimation ($\alpha = \hat{\alpha}$) in order to maximize the correlation output. This procedure is illustrated in the block diagram of Fig. 5.

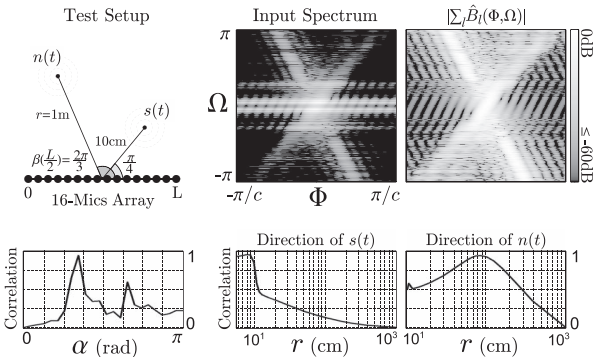

Fig. 6. In this experiment, $s(t)$ and $n(t)$ are positioned such that $\alpha = \left\{\frac{\pi}{2.94}, \frac{2\pi}{3.01}\right\}$ and $r = \{10\text{cm}, 1\text{m}\}$, where SNR $= 5\,\text{dB}$. The angular correlation results in two peaks at $\alpha = \left\{\frac{\pi}{2.88}, \frac{2\pi}{3.07}\right\}$, whereas the radial correlations display a peak at $r = 7.9\text{cm}$ and 0.9m in each direction.

In a multiple source environment, (17) can be used to estimate the number of dominant sources and their respective angles, while (18) is used to estimate the distance of each detected source. A simulation example with a speech source in the near-field plus a white noise interferer in the far-field is shown in Fig. 6.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrate how a spatio-temporal spectral representation of the multichannel input can be used to translate near-field beamforming into a 2D sampling operation and source localization into a parametric estimation problem, due to the inherent property of the representation that separates the signal components from the spatial components in the acoustic wave field. Further work will include the derivation of a more flexible parametric model that takes into account: (i) the imperfections in the microphone positioning and the functional responses, and (ii) the exact convolution result between the spatial window and the near-field energy spreading function, such that the later can be compensated through deconvolution.

6. REFERENCES

[1] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, 1972.

[2] B. Widrow, K. Duvall, R. Gooch, and W. Newman, "Signal cancellation phenomena in adaptive antennas: Causes and cures," *IEEE Trans. Antennas and Propagation*, vol. 30, pp. 469–478, 1982.

[3] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. 30, pp. 27–34, 1982.

[4] F. Pinto and M. Vetterli, "Wave field coding in the spacetime frequency domain," in *IEEE Inter. Conf. Acoustics, Speech and Signal Processing*, 2008.

[5] E. Williams, *Fourier Acoustics*. Academic Press, 1999.

[6] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Processing*, vol. 54, pp. 3790–3804, 2006.